WORD-LEVEL EMPHASIS MODELLING IN HMM-BASED SPEECH SYNTHESIS

K. Yu, F. Mairesse and S. Young

Cambridge University Engineering Department, Trumpington Street, Cambridge, CB2 1PZ, UK
Email: {ky219, f.mairesse, sjy}@eng.cam.ac.uk

ABSTRACT

Expressive speech synthesis has recently attracted great interest. Word-level emphasis is an important form of expressiveness to distinguish between what is the focus of the utterance, and what the computer system expects to be known by the user. Previous work on emphasis synthesis requires emphatic data collected specifically for that task. In this paper, a statistical approach that models and extracts word-level emphasis patterns from *natural* speech is investigated within the HMM based speech synthesis framework. Compared to emphatic speech collected specifically for this task, the cues of emphasis in natural speech are weaker and heavily affected by various suprasegmental features. Two new decision tree clustering approaches, two-pass and factorized decision tree, are proposed to effectively address this problem. Experiments show that both approaches can convey emphasis significantly better than traditional decision tree clustering and HMM adaptation. While the two-pass decision tree approach outperformed the factorized decision tree approach in an emphasis synthesis test, the latter led to significantly better naturalness and hence achieved a better overall balance.

Index Terms— HMM based synthesis, expressive speech synthesis, decision tree

1. INTRODUCTION

The intelligibility of synthesized speech has greatly improved over the past years. Recently, the expressiveness of synthesized speech has been attracting more interest, especially in the area of intelligent spoken dialogue system (SDS). Dialogue system utterances typically combine a large range of information, such as answers to specific user queries, implicit or explicit confirmations, as well as grounding information. Word-level *emphasis* is an important form of expressiveness in these kinds of scenarios. It provides a natural and concise way to distinguish between what is the focus of the utterance, and what the system expects to be known by the user. Disambiguating what is in focus is an important capability for SDSs, as it clarifies the beliefs of the system, and it can affect the expected user response (e.g. 'do you want a cheap *restaurant*?' and 'do you want a *cheap* restaurant?'). Recent research also shows that appropriate emphasis improves the overall perception of synthesized speech [1].

Emphasis synthesis has been mostly investigated within the unit selection framework. A typical approach is to use handcrafted rules to modify the F0 contour [2] or concatenate units from an emphatic corpus [3]. While unit selection methods produce high quality speech, they lack flexibility regarding expressive variations such as emphasis, as the unit coverage required for modelling expressive behaviour quickly becomes prohibitive. On the other hand, HMM-based synthesis (HTS) provides a data-driven framework that allows finer-grained control of the system's expressiveness, by

learning models mapping supra-segmental context features to individual speech parameters. Some recent work shows that HMM based synthesis can produce recognizable variation when modelling emphasis of contrastive words [4]. However, previous work has relied on emphatic data collected specifically for this task, in which emphasis information is predetermined at the discourse or utterance level. Though this type of data gives clear and well defined expressiveness variation, the collection can be very costly for multiple types of variations.

In this paper, statistical approaches to model and extract *word-level* emphasis patterns from *natural* speech with no explicit emphasis variation are investigated within the HMM based statistical speech synthesis (HTS) framework. Compared to emphatic speech data collected specifically for this task, emphasis cues found in *natural* speech are more vague and heavily affected by suprasegmental features. Consequently, traditional decision tree based state-clustering may not be effective to capture the emphasis context features. A common approach to achieve improvement is to construct an emphasis-independent model and adapt the model to speech with emphasis. However, due to the very limited amount of the emphasized words and roughness in emphasis annotation in natural speech, the common adaptation approach may not work well, either. In this paper, two new approaches are investigated to address the problem. In the first method, the state-clustering process is divided into two passes. The first pass constructs an initial decision tree by using only emphasis context features. In the second pass, each leaf node of the initial decision tree is further split using the normal question set to form the final decision tree. Though this approach effectively prioritizes emphasis questions, it fragments the training data and reduces the amount of data that can be used by the second pass decision tree clustering. A second approach, factorized decision tree, is then proposed to address the problem. It uses two sets of parameters associated with two decision trees to represent the emphasis and non-emphasis factors respectively. A set of canonical Gaussian parameters represent normal speech variation and are associated with the decision tree constructed using the normal question set. A set of linear transforms represent the relationship between emphasized and non-emphasized parameters and are associated with the decision tree constructed using emphasis-related questions. The final HMM parameters are shared for each intersection of the two decision trees and can be constructed by applying the emphasis transform to the canonical Gaussian parameters. With this factorized representation and interleaving update of both sets of parameters, the emphasis information can be effectively modelled without seriously affecting normal speech variability.

The rest of this paper is structured as follows. Section 2 describes the approaches used for word-level emphasis modelling, including the traditional decision tree clustering and adaptation methods, as well as two new approaches. Section 3 gives details of the emphasis information extraction from natural speech. Subjective

tests for measuring the ability to convey emphasis and the overall naturalness are described in section 4, followed by the conclusion.

2. WORD-LEVEL EMPHASIS MODELLING APPROACHES

To model word-level emphasis in natural speech, we rely upon training data labelled with emphasized words or phrases. These labels are then used as additional context information during the training of HMMs. During synthesis, emphasis labels are explicitly generated for the words to be emphasized. The word-level emphasis modelling approaches differ in how this additional emphasis context information is used.

As there are a large number of possible contexts, decision tree based state clustering with the minimum description length (MDL) criterion is commonly used for robust parameter estimation. Spectrum and F0 are normally modelled in separate streams with different decision trees. Unless noted otherwise, multi-space density HMMs (MSDHMM) are used in this paper to model the F0 stream [5].

2.1. Decision tree with additional emphasis context questions

In this paper, the emphasis information is represented by three additional emphasis-related supra-segmental context features, i.e. characterizing whether the current, previous and following words are emphasized. A straightforward way to incorporate these features is to introduce an emphasis-related decision tree question for each possible combination of emphasis context features (e.g., are the previous and the current words emphasized?) and add them to the normal context question set. Supra-segmental context features have also been used for expressive synthesis in previous work [6, 4]. At synthesis time, the emphasis-dependent questions in the final decision tree determine the final model being used for synthesizing emphasis.

The normal context features include phones, counts, positions, syllables, words and phrases, lexical stress and pitch accents, etc. In contrast to emphatic speech collected specifically for this task, in natural speech emphasis context features usually have less clustering power than normal context features. This could potentially limit the number of emphasis context questions to be asked during decision tree clustering.

2.2. Emphasis adaptation

A commonly used approach to capture specific context features with a small amount of data is to use adaptation techniques. In this paper, to get more focus on the emphasis features, we apply standard maximum likelihood linear regression (MLLR) [7] to synthesize emphasis. As previous methods on adaptation for expressive TTS typically focus on utterance-level variation [6], they do not evaluate whether the boundaries between different stylistic effects are modelled correctly. In this paper, we first partition the data into emphasis and non-emphasis regions according to the word/phrase boundaries. Then, we train emphasis-independent HMMs by ignoring emphasis labels in the dataset, and learn two sets of mean MLLR transforms that adapt the HMMs to the emphasized and non-emphasized region respectively. A regression class tree is used to generate multiple mean MLLR transforms. This is similar to the standard adaptation technique except that the supervision data consists of word/phrase segments rather than complete utterances.

At synthesis time, emphasis transforms are applied to HMMs belonging to emphasized phrases, and non-emphasis transforms are applied to the remaining words. The speech parameters are then generated from the sequence of adapted HMMs.

2.3. Two-pass decision tree construction

As indicated in section 2.1, directly putting emphasis context questions into the question set may have little effect. In order to address this issue, we present a two-pass decision tree state clustering approach that computes a first decision tree from emphasis-related context features only, and then extends that tree in a second pass with the full set of normal context features.[1] Figure 1 illustrates the outcome of the two passes. This method effectively forces emphasis-related features to be at the top of the tree based on their clustering power, thus trading off suboptimal parameter tying for additional expressiveness (the number of models increases by 33.9% compared with the regular state clustering approach).
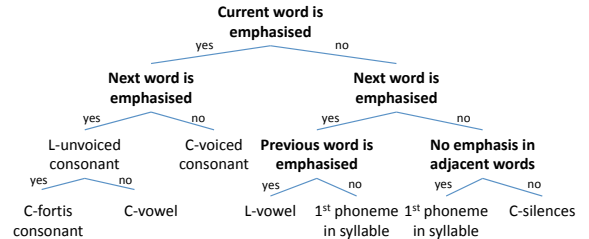

Fig. 1. Partial state clustering decision tree resulting from the two-pass extension (log F0). The nodes of the tree produced during the first pass are in bold (C/L/R = current/left/right segment).

2.4. Factorized decision tree

Although the two-pass decision tree approach can effectively exploit emphasis questions, it fragments the training data and leads to a reduction of the amount of the data that can be used in state clustering with normal context questions. The larger the set of emphasis questions, the more serious the training data sparsity problem will be. Hence, the focus on emphasis is achieved at the cost of reducing the amount of effective training data. To address this problem, a *factorized decision tree* approach is proposed.
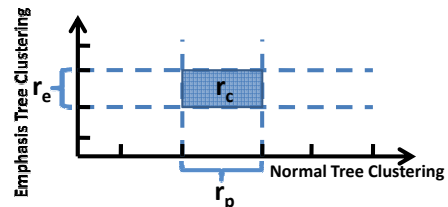

Fig. 2. Combination of Phone/Position and Emphasis Decision Trees

During acoustic realization of phones in context, there are many factors which may affect the process. The prior knowledge of those factors form the questions used in the decision tree based state clustering procedure. Due to the nature of the factors, some questions are highly correlated, for example, the phonetic broad class questions and the syllable questions. However, other questions are relatively weakly correlated, such as the phonetic broad class questions and the emphasis questions. To reflect the independence between factors and reduce the data sparsity problem, it is useful to introduce some factorized form of model representation. Therefore, rather than pooling all questions to construct a single decision tree, two decision trees are built, each corresponding to different factors. One decision tree is constructed using only the normal context questions (e.g. phone and position questions) while the other one is constructed with emphasis related questions [2]. After the two decision trees are built sep-

---

[1] The two-pass extension method has previously been used for unsupervised speaker adaptation [8].

[2] In this paper, each emphasis related question consists of one emphasis context feature and one normal context feature. This will lead to powerful transforms as the number of transforms is large.

arately, the emphasis decision tree is appended to each leaf node of the normal decision tree to get further split clusters as shown in Fig 2.

The leaf nodes of the combined decision tree, r_c correspond to the intersections of the leaf nodes of the emphasis decision tree r_e and the normal decision tree r_p. Hence, r_c are atomic leaf nodes which can construct r_e and r_p. Assuming there are N_e and N_p leaf nodes in the emphasis tree and phonetic tree respectively, the combined decision tree could have at most $N_e \times N_p$ leaf nodes. The state output distribution parameters within r_c are tied and represented in a factorized form to reflect the nature of the two trees. The factorization approach adopted is to use canonical model parameters to represent normal context factors, and to use linear transforms to represent the emphasis factor. When single Gaussian distributions are used, the mean vector of the combined leaf node is represented by

$$\boldsymbol{\mu}^m = \boldsymbol{\mu}^{r_c} = \mathbf{A}^{r_e}\boldsymbol{\mu}^{r_p} + \mathbf{b}^{r_e} = \mathbf{W}^{r_e}\boldsymbol{\xi}^{r_p} \quad (1)$$

where m is the index of the Gaussian distribution associated with the atomic leaf node r_c, which is the intersection of the leaf node in emphasis decision tree, r_e, and the leaf node of phonetic decision tree r_p. $\boldsymbol{\xi}^{r_p} = [\boldsymbol{\mu}^{r_p T}\ 1]^T$ is the generalized mean vector of leaf node r_p while $\mathbf{W}^{r_e} = [\mathbf{A}^{r_e}\ \mathbf{b}^{r_e}]$ is the transform associated with leaf node r_e. From equation (1), the parameters of the combined leaf node can not be directly estimated. Instead, they are constructed using two sets of parameters with different state clustering structures. With this factorized representation, the estimation of the transform parameters for cluster r_e and Gaussian parameters for cluster r_p has to be interleaved. The detailed procedure is as follows:

1. Get initial parameters of $\boldsymbol{\mu}^{r_p}$ from state clustering using normal decision tree and let $\boldsymbol{\mu}^m = \boldsymbol{\mu}^{r_p}, m \in r_p$.

2. Estimate $\mathbf{W}^{r_e}$ given the current model parameters. This is a standard MLLR estimate [7]. The dth row is estimated as

$$\mathbf{w}_d^{r_e} = \mathbf{G}_d^{-1}\mathbf{k}_d \quad (2)$$

where the sufficient statistics for the dth row are given by

$$\mathbf{G}_d = \sum_t \sum_{m \in r_e} \frac{\gamma_m(t)}{\sigma_{dd}^{r_p(m)}}\boldsymbol{\xi}^{r_p(m)}\boldsymbol{\xi}^{r_p(m)T} \quad (3)$$

$$\mathbf{k}_d = \sum_t \sum_{m \in r_e} \frac{\gamma_m(t)o_{t,d}}{\sigma_{dd}^{r_p(m)}}\boldsymbol{\xi}^{r_p(m)} \quad (4)$$

where $o_{t,d}$ is the dth element of observation vector $\mathbf{o}_t$, $\sigma_{dd}^{r_p(m)}$ is the dth diagonal element of $\boldsymbol{\Sigma}^{r_p(m)}$, $r_p(m)$ is the leaf node of the normal decision tree to which Gaussian component m belongs, $\gamma_m(t)$ is the posterior for Gaussian component m at time t which is calculated using forward-backward algorithm with the parameters from equation (1).

3. Estimate $\boldsymbol{\mu}^{r_p}$ given the emphasis transform parameters. This is similar to the mean update in speaker adaptive training [9]. Given the sufficient statistics

$$\mathbf{G} = \sum_t \sum_{m \in r_p} \gamma_m(t)\mathbf{A}^{r_e(m)T}\boldsymbol{\Sigma}^{r_p(m)-1}\mathbf{A}^{r_e(m)}$$

$$\mathbf{k} = \sum_t \sum_{m \in r_p} \gamma_m(t)\mathbf{A}^{r_e(m)T}\boldsymbol{\Sigma}^{r_p(m)-1}\left(\mathbf{o}_t - \mathbf{b}^{r_e(m)}\right)$$

the new mean is estimated by

$$\boldsymbol{\mu}^{r_p} = \mathbf{G}^{-1}\mathbf{k} \quad (5)$$

where $r_e(m)$ is the leaf node of emphasis tree that Gaussian component m belongs to.

4. Given the updated mean and transform, the re-estimation of $\boldsymbol{\Sigma}^{r_p}$ is performed using the standard covariance update formula except that the statistics are accumulated for each leaf node r_p rather than the atomic leaf node r_c.

As in HMM based synthesis, spectrum and F0 are normally modelled using separate streams, the factorized decision trees are also built for each stream respectively. Note that the factorized representation of atomic leaf node parameters can have various forms. For example, when using multiple decision trees for F0 modelling in [10], multiple-cluster mean vectors and interpolation weights were used to construct the mean vector of the atomic leaf node.

3. DATA PREPARATION

The emphasis modelling techniques described in the previous section were evaluated in a *natural emphasis* synthesis task. The training data is a subset of the male English voice with a Scottish accent (awb) in the CMU ARCTIC speech database. One judge annotated the 597 utterances of the set A of the dataset, by labelling the word(s) that were perceived as the focus of the utterance based on the natural emphasis of the speaker.[3] It is worth noting that there was no intention of collecting speech with emphasis during the construction of the ARCTIC speech database, hence, it does not contain strong stylistic variation. The emphasis labels were given to the naturally emphasized words (e.g., content words) as well as involuntary fluctuations of the speaker. The judge labelled 2.32 emphasized words per utterance on average (26.3% of the words). In order to assess the reliability of the annotation, a second judge annotated a subset of 50 utterances of 597 sentences. This yielded an agreement of 1.04 words per utterance on average, and a disagreement of 1.52 words on average.[4] This suggests that the natural emphasis information obtained from a human judge is highly subjective. However, most of the disagreements were due to a difference of granularity when labelling emphasis, as there was an overlap for 72% of the utterances. This shows that there exists consistent *rough agreement* over natural emphasis. Though emphasis is likely to be harder to capture when it is not explicitly generated in natural speech, techniques that can extract the emphasis component from natural data can significantly reduce the cost of stylistic modelling.

4. EXPERIMENTAL RESULTS

All systems were built using a modified version of HTS HMM-synthesis toolkit version 2.1. Three systems (direct emphasis context decision tree, emphasis adaptation and two-pass decision tree) were built using MSDHMM. The factorized decision tree system was built using HMM-GTD [11] because it is easier to apply full transforms to static and dynamic F0 features with HMM-GTD. In all experiments, 6 emphasis context questions were used. The static feature set comprised 25 Mel-Cepstral coefficients, log F0 and aperiodic energy components in five frequency bands (0 to 1, 1 to 2, 2 to 4, 4 to 6 and 6 to 8 KHz). All features were extracted using the STRAIGHT speech analysis system. During HMM training, the stream weight for the aperiodic component was set to zero. Hence, the forward-backward alignments depended only on the spectral and F0 features. Statistics for the aperiodic components were however collected and their parameters were updated in the standard way.

First, a subjective test was performed to measure the ability to convey emphasis. For each system, 10 utterances generated with and without emphasis were provided to each listener for pair-wise comparison. When perceiving a difference of emphasis, the listener was

---

[3]Available at http://mi.eng.cam.ac.uk/~farm2/emphasis.
[4]Cohen's Kappa cannot be used here because the phrases are not distinct elements.

asked to select the word that carried the additional emphasis. Each listener evaluated a randomized set of 40 utterances in the tourist information domain, with one emphasized word in each utterance (e.g., 'Char Sue is an *expensive* Chinese restaurant'). Altogether 23 listeners, 11 native and 12 non-native, participated in the test. The average number of emphasized words conveyed correctly is shown in Fig. 3.
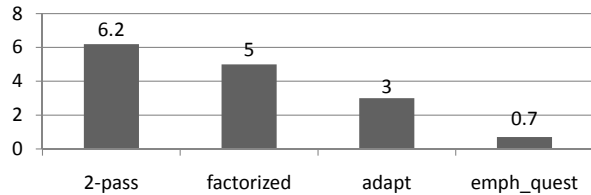

Fig. 3. Emphasis synthesis results for different systems.

It can be observed that two-pass decision tree best conveyed emphasis, while the direct emphasis context decision tree almost did not convey any emphasis. The difference between any two systems is significant at a p value of 0.01.

By checking the questions asked during decision tree clustering, it was found that emphasis questions were rarely asked when using the direct emphasis context approach, which results in limited perceptible difference between emphasized and non-emphasized speech. This is likely to be due to (a) the poor clustering power of the emphasis context features compared with normal context features, (b) their correlation with other features (e.g. content words or accent features) [5] as well as (c) the lack of emphasis variation in the annotated natural speech. Using the emphasis adaptation approach improved emphasis conveying. However, as the regression base classes were built on the states clustered using phonetic and position questions and the aligned word boundaries may not be accurate, the absolute performance is still very low. By using the factorized representation, the emphasis questions were all effectively used in HMM training and the emphasis synthesis performance was significantly improved. About half of the emphasized words were correctly identified. The two-pass decision tree approach forces emphasis questions to have the highest priority during state-clustering, hence led to the most perceptible emphasis.
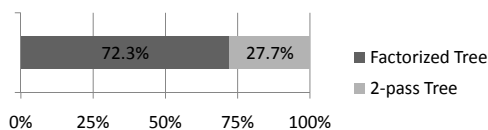

Fig. 4. Subjective comparison between 2-pass and factorized decision tree approaches

Although the two-pass decision tree approach conveyed emphasis the most successfully, the additional number of model parameters can potentially affect naturalness. To evaluate this claim, a preference choice test on 10 sentences were performed to compare the two-pass decision tree approach to the factorized decision tree approach. 5 of the 10 sentences were with emphasized data and the other 5 without. Two wave files were synthesized for each sentence. To reduce the variance introduced by forcing the user to make a choice, the 10 wave file pairs were duplicated and the order of the two systems were swapped. The final 20 samples were then shuffled and provided to the listeners. Each listener was asked to select the more

---

[5]Controlling accent features directly was unsuccessful, as such features are automatically generated from the transcription, thus ignoring the acoustic data.

natural example from each wave file pair. Altogether 16 listeners, 8 native and 8 non-native, participated in the listening test. The result is shown in Fig. 4.

Statistical significance tests were performed assuming a binomial distribution of each choice. It was found that the factorized decision tree approach was perceived as significantly more natural than the two-pass decision tree approach. This shows that the factorized decision tree approach can yield a better balance between conveying emphasis and maintaining a high naturalness.

5. CONCLUSION

Word-level emphasis is an important form of expressiveness. Previous work has focused on using emphatic speech data collected specifically to model emphasis. This paper investigates word-level emphasis modelling techniques for natural speech. Due to the weakness of emphasis cues, directly using emphasis context features and the traditional adaptation approach does not work well. A two-pass decision tree and a factorized decision tree approach are proposed to address this problem. Experiments showed that the two-pass decision tree is most effective at conveying emphasis while the factorized decision tree approach provides a good balance between conveying emphasis and maintaining naturalness.

6. REFERENCES

[1] V. Strom, A. Nenkova, R. Clark, Y. Vazquez-Alvarez, J. Brenier, S. King, and D. Jurafsky, "Modelling prominence and emphasis improves unit-selection synthesis," in *Proceedings of Interspeech*, 2007.

[2] "The Festival speech synthesis system," http://www.cstr.ed.ac.uk/projects/festival/.

[3] A. Raux and A. Black, "A unit selection approach to F0 modeling and its application to emphasis," in *Proc. ASRU*, 2003.

[4] L. Badino, J. S. Andersson, J. Yamagishi, and R. Clark, "Identification of contrast and its emphatic realization in hmm based speech synthesis," in *Proc. INTERSPEECH*, 2009.

[5] K. Tokuda, T. Mausko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D, no. 3, pp. 455–464, 2002.

[6] Junichi Yamagishi, Koji Onishi, Takashi Masuko, and Takao Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis," *IEICE TRANSACTIONS on Information and Systems*, vol. E88-D, no. 3, pp. 502–509, 2005.

[7] C. J. Leggetter and P. C. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density HMMs," *Computer Speech and Language*, vol. 9, pp. 171–186, 1995.

[8] Matthew Gibson, "Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models," in *Proc. INTERSPEECH*, 2009.

[9] T. Anastasakos, J. Mcdonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proc. ICSLP*, 1996, pp. 1137–1140.

[10] H. Zen and N. Braunschweiler, "Context-dependent additive log F0 model for HMM-based speech synthesis," in *Proc. INTERSPEECH*, 2009.

[11] K. Yu, T. Toda, M. Gasic, S. Keizer, F. Mairesse, B. Thomson, and S. Young, "Probablistic modelling of F0 in unvoiced regions in HMM based speech synthesis," in *Proc. ICASSP*, 2009.